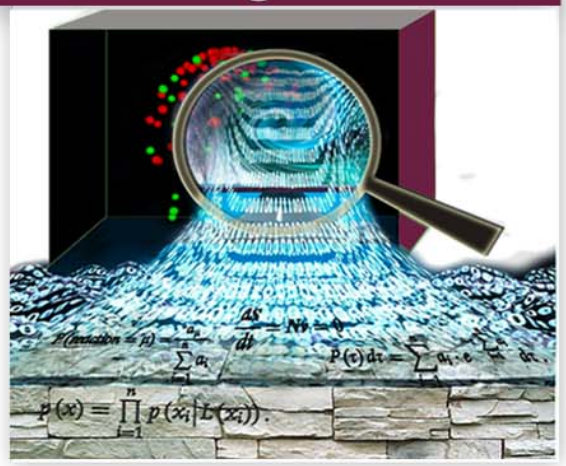
U.S. Department of Energy's Office of Science

From Data to DISCOVERY

Dr. Raymond L. Orbach
Under Secretary for Science
U.S. Department of Energy
www.science.doe.gov

Finding the Dots, Connecting the Dots, Understanding the Dots

Presented to
2007 DICE Alliance Meeting
May 7-9, 2006
Springfield, Ohio



DOE Office of Science

Supports basic research that underpins DOE missions. Constructs and operates large scientific facilities for the U.S. scientific community (accelerators, synchrotron light sources, neutron sources, supercomputers).

Seven Program Offices

- Advanced Scientific Computing Research (ASCR)
- Basic Energy Sciences (BES)
- Biological and Environmental Research (BER)
- Fusion Energy Sciences (FES)
- High Energy Physics (HEP)
- Nuclear Physics (NP)
- Workforce Development (WD)

Supports basic research that underpins DOE missions at

- Seventeen National Laboratories
- 280 Research Universities

Facilities for the Future of Science
A Twenty Year Outlook

- 28 Prioritized Projects for the near-term, mid-term, and long-term

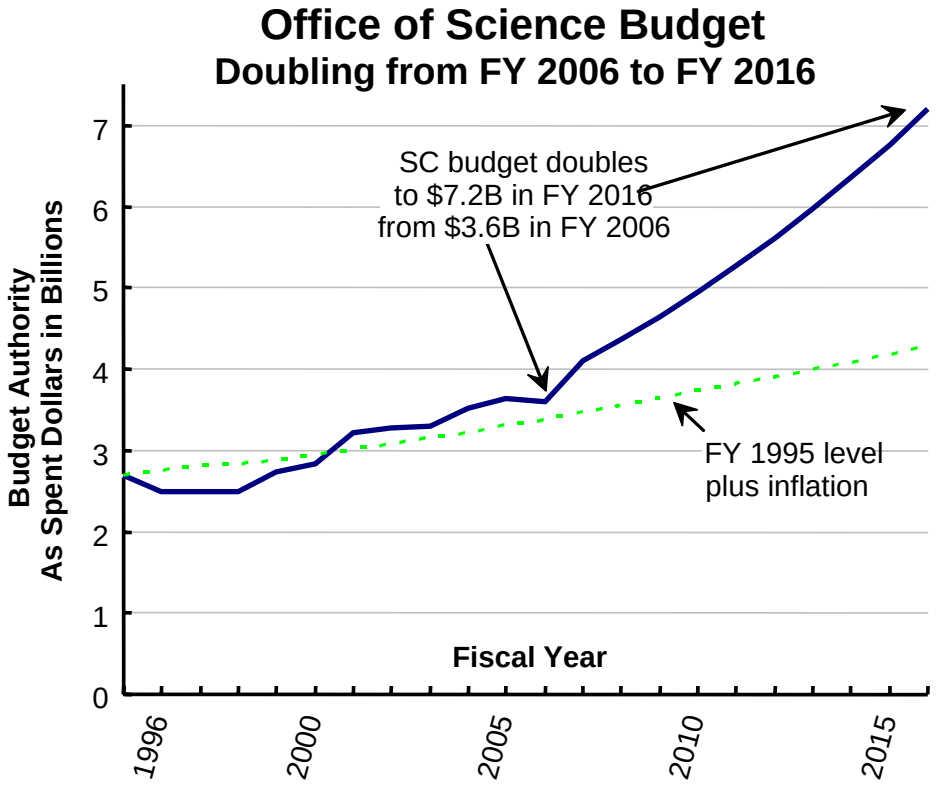
1. ITER
2. Ultrascale Scientific Computing Capability (*Implemented as Leadership Computing Facilities (LCFs)*)
3. *Joint Dark Energy Mission*
3. *Linac Coherent Light Source*
3. *Protein Production and Tags (Implemented as Bioenergy Centers)*
3. *Rare Isotope Beam Facility*
7. *Characterizing and Imaging (Implemented as Bioenergy Centers)*
7. *CEBAF upgrade*
7. *ESnet upgrade*
7. *NERSC upgrade*
7. *Transmission Electron Achromatic Microscope*
12. *BTeV (Terminated)*
13. *Linear Collider*
14. *Analysis/Modeling of Celular Systems (Implemented as Bioenergy Centers)*
14. *SNS 2-4 Megawatt Upgrade*
14. *SNS Second Target Station*
14 *Whole Proteome Analysis (Implemented as Bioenergy Centers)*
18. *Double Beta Decay Underground Detector*
18. *Next-Step Spherical Torus*
18. *RHIC II*
21. *National Synchrotron Light Source Upgrade (technology readiness change)*
21. *Super Neutrino Beam*
23. *Advanced Light Source Upgrade*
23. *Advanced Photon Source Upgrade*
23. *eRHIC, eLIC, or the Electron Ion Collider*
23. *Fusion Energy Contingency*
23. *HFIR Second Cold Source and Guide Hall*
23. *Integrated Beam-High Energy Density Physics Experiment*

www.science.doe.gov/Scientific_User_Facilities/History/20-Year-Outlook-screen.pdf

Funding Umbrella:
American Competitiveness Initiative

"First, I propose to double the federal commitment to the most critical basic research programs in the physical sciences over the next 10 years. This funding will support the work of America's most creative minds as they explore promising areas such as nanotechnology, supercomputing, and alternative energy sources."

President George W. Bush
State of the Union Address, Jan. 31, 2006



- The FY 2007 President's request for science funding is a **14.1% increase** and sets the Office of Science on a path to **doubling by 2016**.
- An historic opportunity for our country – a renaissance for U.S. science and continued global competitiveness.

Status of Near-term Facilities in 20-Year Outlook
By the end of FY 2007

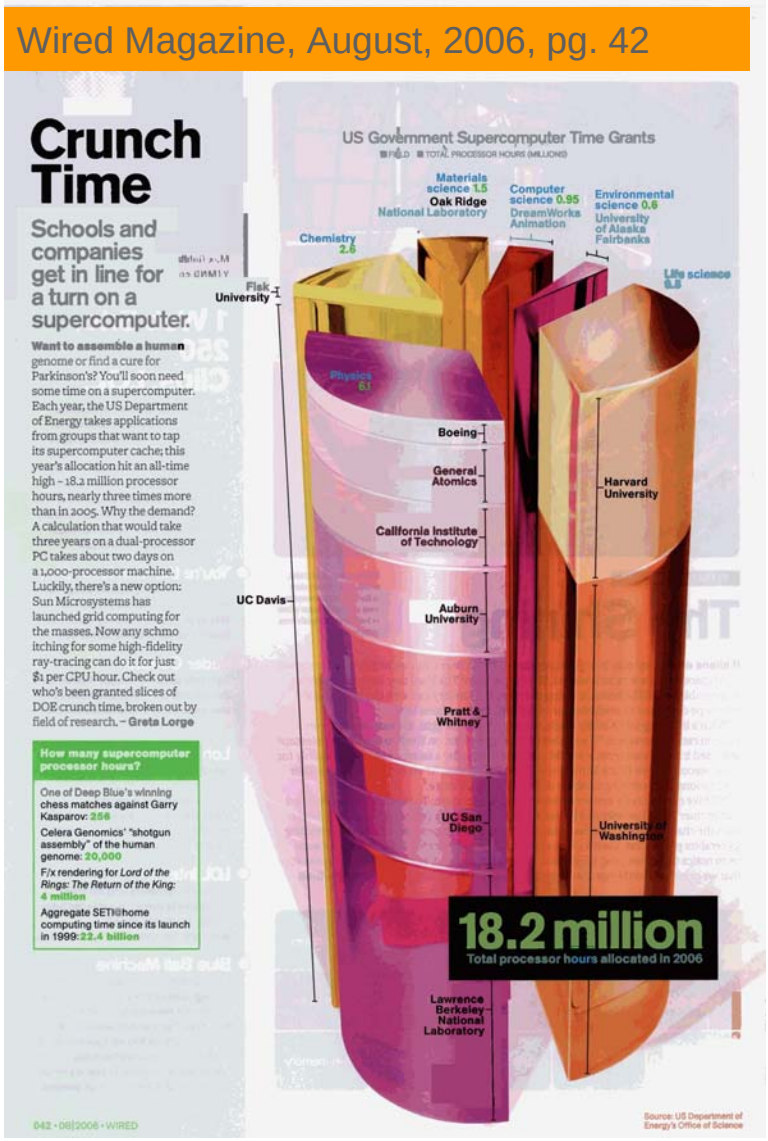
Near-Term

Priority	Program	Facility		R&D	Conceptual Design	Engineering Design	Construction	Operation
1	FES	ITER						
2	ASCR	UltraScale Scientific Computing Capability						
Tie for 3	HEP	Joint Dark Energy Mission						
	BES	Linac Coherent Light Source						
	BER	Protein Production and Tags →	Bioenergy Research Centers*					
	NP	Rare Isotope Beam Facility (previously RIA) #						
Tie for 7	BER	Characterization and Imaging →	Bioenergy Research Centers*					
	NP	CEBAF Upgrade						
	ASCR	ESnet Upgrade						
	ASCR	NERSC Upgrade						
	BES	Transmission Electron Achromatic Microscope						

100% complete 50-75% complete 25-50% complete 1-25% complete

Facilities for the Future of Science
Ultrascale Scientific Computing

Wired Magazine, August, 2006, pg. 42

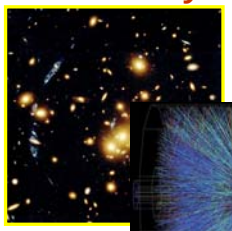


- LCF at Oak Ridge (Capability)
  - Cray XT system, currently at 119 teraflops, planned for 1 petaflop by the end of 2008.
- Argonne LCF (Capability)
  - IBM Blue Gene system, currently at 5.7 teraflops, planned for 250-500 teraflops in 2008.
- NERSC (Capacity) at Lawrence Berkeley
  - Currently at 10 teraflops, planned for 100-150 teraflops by the end of 2008 – serving ~2,000 users.
- Environmental and Molecular Science Lab at Pacific Northwest
  - HP system, currently at 11.5 teraflops, planned for 100 teraflops by the end of 2008.
- Innovative and Novel Computational Impact on Theory and Experiment (INCITE)
  - Computing time and technical support to a limited number of large scale computational projects promising breakthroughs in science and engineering.
  - Peer reviewed process open to all.
  - 18.2 million CPU hours to 15 projects in 2006.
  - 95 million CPU hours to 45 project in 2007.
  - 250 million CPU hours for 2008.

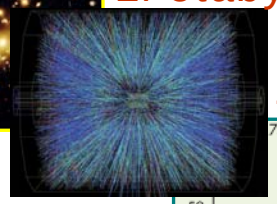
Scientific Discovery through Advanced Scientific Computing

Three Pillars of Scientific Discovery: Experiment, Theory, and Simulation

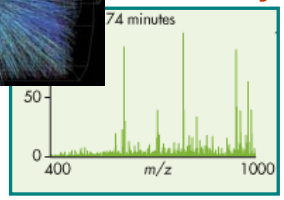
10Terabytes/day



2Petabytes/exp



1Petabyte/yr

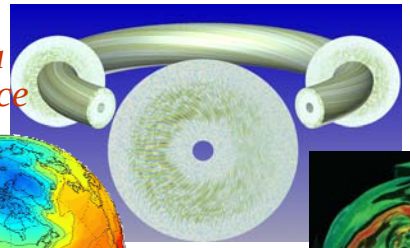


Astro-physics

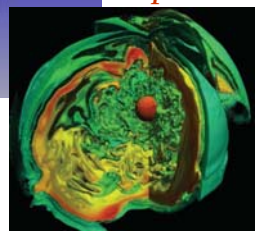
Nuclear Physics (Star)

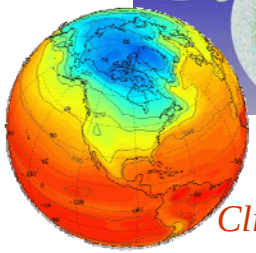
Biology (Peptide data)

Fusion Plasma turbulence



Supernova





Climate

50-500TB/simulation

Two different kinds of very large data sets:

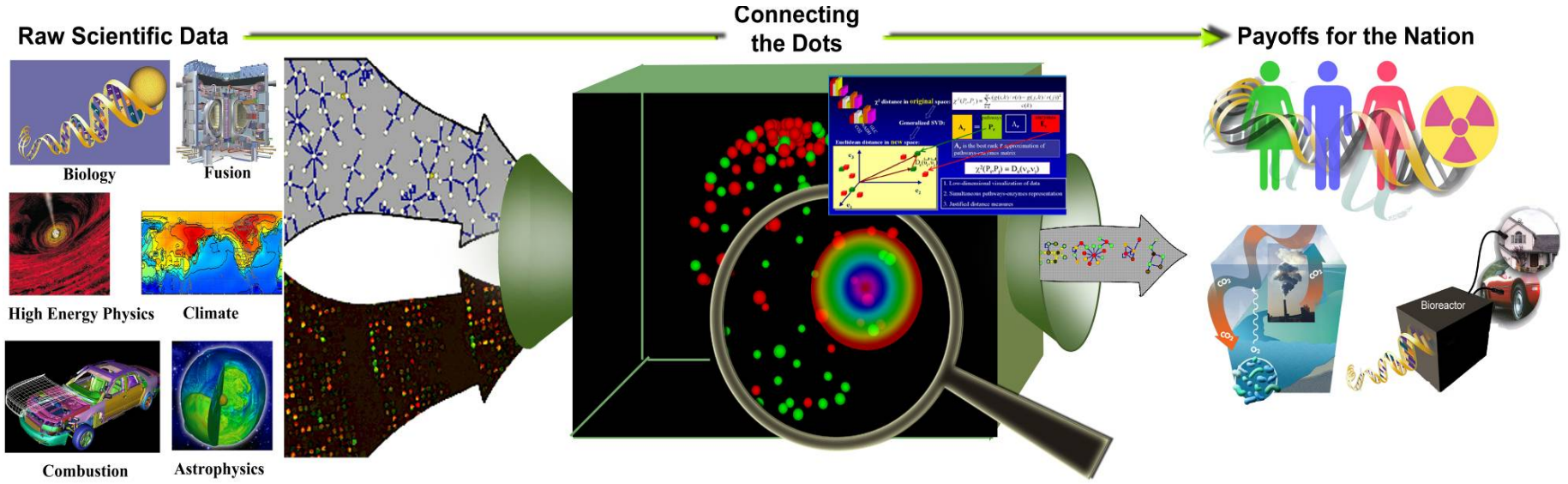
Experimental data

- High energy physics, power grids, environment and climate observation data, cosmology, biological mass-spectrometry.
- Data needs to be retained for long term.

Simulation data

- Nuclear energy, astrophysics, climate, fusion, catalysis, Lattice Quantum Chromodynamics.
- From computationally intensive simulations.
- Post processing of data using quantum Monte Carlo, clustering, Single Value Decomposition, perturbation theory, and molecular dynamics.

Scientific Discovery



Sheer Volume of Data

Climate

Now: 20-40 Terabytes/year

5 years: 5-10 Petabytes/year

Fusion

Now: 100 Megabytes/15 min

5 years: 1000 Megabytes/2 min

Advanced Mathematics and Algorithms

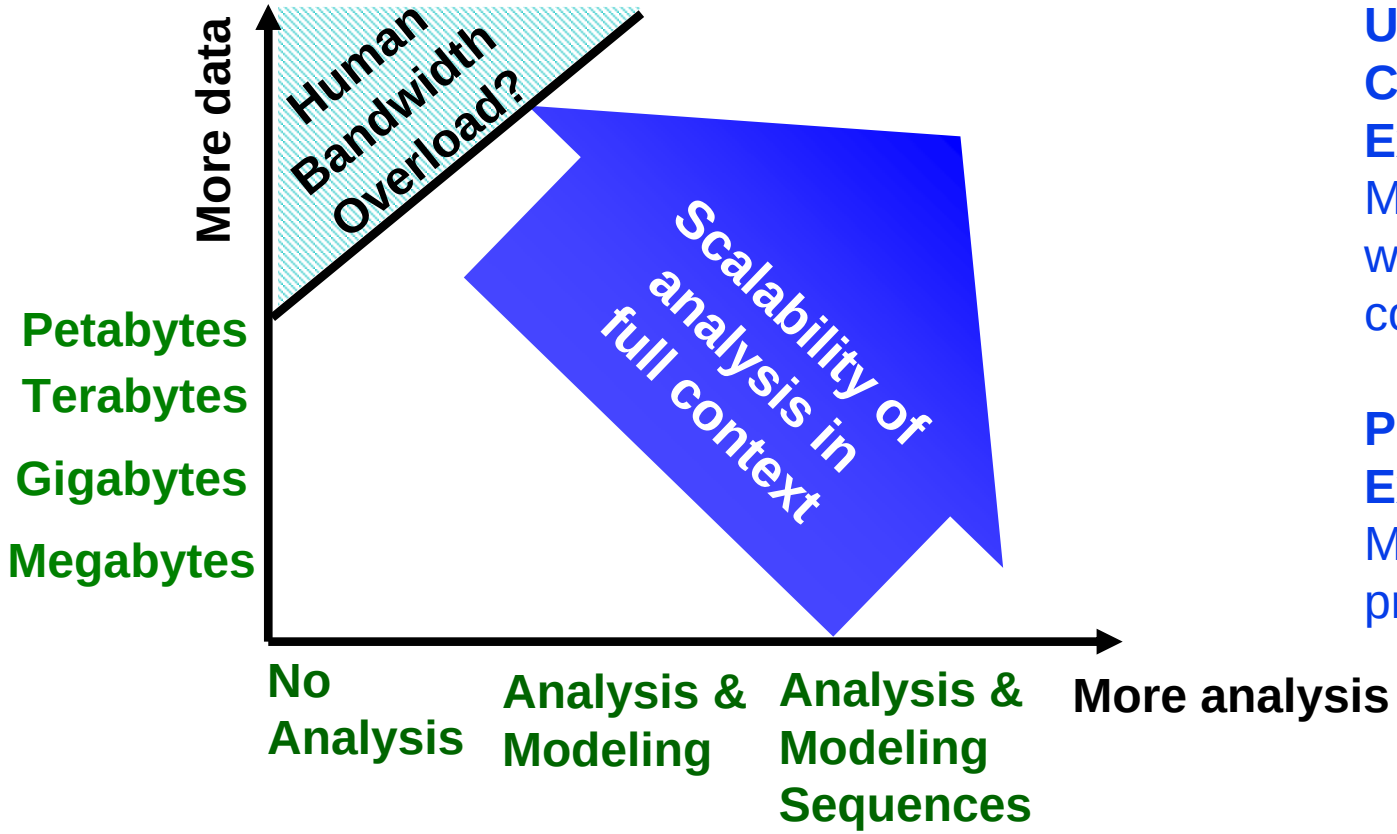
- Huge dimensional space
- Combinatorial challenge
- Complicated by noisy data
- Requires high-performance computers

Providing Predictive Understanding

- Produce hydrogen-based energy
- Stabilize carbon dioxide
- Clean and dispose toxic waste

Making Sense of Data

Not humanly possible to browse a petabyte of data. Analysis must select views or reduce to quantities of interest.



Ultrascale Computational Experiment:
Must be smart about which probe combinations to see!

Physical Experiment:
Must be smart about probe placement!

To see 1 percent of a petabyte at 10 megabytes per sec. takes 35 8-hour days!

Large Noisy Data Sets:
Optimal Set of Data Clusters

Mathematical and Computational Challenges and Needs

- Dimensionality – Interpretation and indexing of high dimensional data.
- Computational Intractability – Naïve approaches to connecting-the-dots problems scale exponentially with the size of the problem.
- What is Noise? – Finding the order in chaos; extracting the signal in noisy data.
- Discovery of the Exemplars – storage of compressed information - optimally condensed description.
  - Example: in the case of gene expression with tens of thousands of sequences, the clusters would be groups of genes with similar patterns of expression - the 'Exemplars'

Where are the “Exemplars?”

Messages
are exchanged in
the directions of
the fingers and of
the glances,
leading to the
recognition of
San Matteo
as the “exemplar.”

Marc Mézard

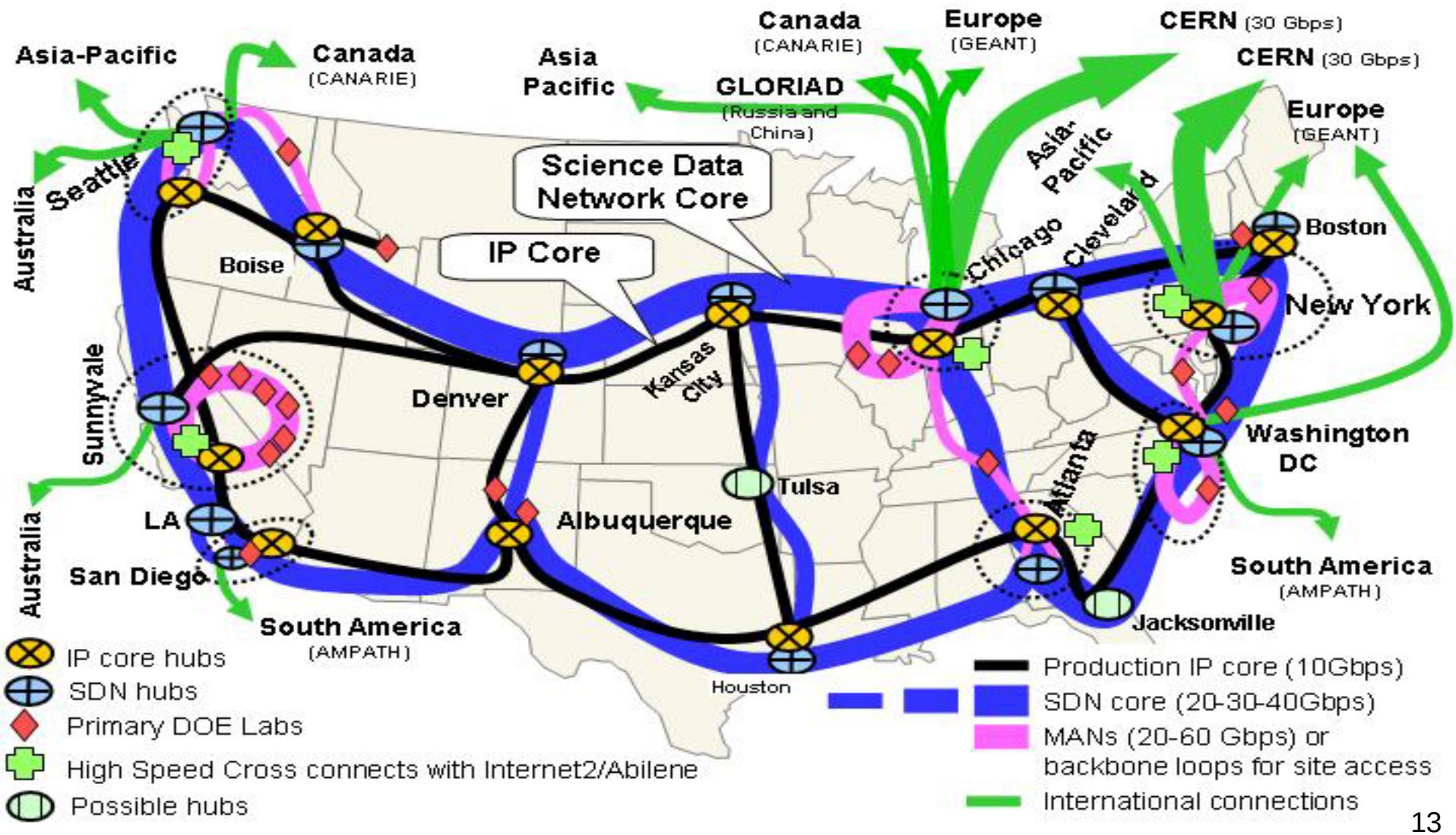
Caravaggio’s
Vocazione di San Matteo
(St. Matthew called to the apostolate)

Challenges in Bringing the Data to the User

Science Drivers Science Areas / Facilities	End2End Reliability	Connectivity	Today End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
Magnetic Fusion Energy	99.999% (Impossible without full redundancy)	• DOE sites; US Universities; Industry	200+ Megabits per second	1 Gigabits per second	• Bulk data • Remote control	• Guaranteed bandwidth • Guaranteed Quality of Service • Deadline scheduling
NERSC and ACLF	-	• DOE sites; US Universities; International; Other Office of Science supercomputers	10 Gigabits per second	20 to 40 Gigabits per second	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Guaranteed Quality of Service • Deadline Scheduling • Public Key Infrastructure (security) / Grid technology
Nuclear Physics (Relativistic Heavy Ion Collider)	-	• DOE sites; US Universities; International	12 Gigabits per second	70 Gigabits per second	• Bulk data	•Guaranteed bandwidth • Public Key Infrastructure (security) / Grid technology
Spallation Neutron Source	High (24x7 operation)	• DOE sites	640 Megabits per second	2 Gigabits per second	• Bulk data	
Advanced Light Source	-	•DOE sites; US Universities; Industry	1 Terabytes per day 300 Megabits per second	5 Terabytes per day 1.5 Gigabits per second	•Bulk data •Remote control	•Guaranteed bandwidth • Public Key Infrastructure (security) / Grid technology
Climate Science	-	•DOE sites; US Universities; International	-	5 Petabytes per year 5 Gigabits per second	• Bulk data • Remote control	•Guaranteed bandwidth • Public Key Infrastructure (security) / Grid technology
High Energy Physics (Large Hadron Collider)	99.95+% (Less than 4 hrs/year)	•US Tier1 (FNAL, BNL); US Tier2 (Universities); International (Europe, Canada)	10 Gigabits per second	60 to 80 Gigabits per second (30-40 Gigabits per second per US Tier1 site)	• Bulk data •Coupled data analysis processes	•Guaranteed bandwidth •Traffic isolation • Public Key Infrastructure (security) / Grid technology

Plans for the ESnet

Core Networks: 40-50 Gigabits per second (Gbps) in 2009, 160-400 Gigabits per second in 2011-2012



The Path Forward
Working together to move ahead

- Office of Science has a long history of partnering
  - Across Disciplines; Institutions; and with Industry
- Data challenges are becoming universal
  - Astrophysics to Electricity Grid Modernization to Google
- Current Investments in the Office of Science
  - Scientific Discovery through Advanced Computing
    - SciDAC supports Centers and Institutes that are focused on overcoming the technical barriers to computational science.
    - Current portfolio includes: Data Management; Visualization and Distributed Computing, and an Outreach Center.
  - Computer Science
    - Core research in architectures, languages, operating systems, file systems, compilers, performance tools, data management and viz.
  - Applied Mathematics
    - Research on algorithms and libraries including multiscale math, the mathematics of large datasets, and optimization of complex systems.

Final Thoughts

- We Invite you to participate in our planning for the petascale, exascale and beyond:
  - Three “town hall meetings” on the proposed Simulation and Modeling at the Exascale for Energy Ecological Sustainability and Global Security (E^3SGS) program:
    - Lawrence Berkeley National Laboratory hosted the first meeting April 17-18
      - http://hpcrd.lbl.gov/E3SGS/main.html

      “ The planned petascale computer systems and the potential for exascale systems shortly provide an unprecedented opportunity for science; one that will make it possible to use computation not only as an critical tool along with theory and experiment in understanding the behavior of the fundamental components of nature but also for fundamental discovery and exploration of the behavior of complex systems with billions of components including those involving humans.”
    - Oak Ridge National Laboratory – May 17-18
      - http://computing.ornl.gov/workshops/town_hall/index.shtml
    - Argonne National Laboratory – May 31-June 1
      - https://www.cls.anl.gov/events/workshops/townhall07/index.php

“The purpose of computing is insight, not numbers”
Richard Wesley Hamming